The Report Committee for Swarup Sahoo
Certifies that this is the approved version of the following Report:

Life Cycle Cost Analysis for Cancer Patients at The University of Texas MD Anderson Cancer Center

APPROVED BY
SUPERVISING COMMITTEE:

James Eric Bickel, Supervisor

Jonathan Bard

Life Cycle Cost Analysis for Cancer Patients at The University of Texas MD Anderson Cancer Center

by

Swarup Sahoo

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 2018

Acknowledgements

This work would not have been possible if not for the guidance and support of Dr. J Eric Bickel, my supervisor. I would like to thank Dr. Jonathan bard, for his encouragement and help. I would also like to thank Mr. Connor Burdine and his team at The UT MD Anderson Cancer Center, for providing this opportunity to conduct research on their data and for their invaluable knowledge and guidance. Lastly, I would like to thank Manoj Gedela, for his help and support.

Abstract

Life Cycle Cost Analysis for Cancer Patients at The University of Texas MD Anderson Cancer Center

Swarup Sahoo, M.S.E

The University of Texas at Austin, 2018

Supervisor: James Eric Bickel

Healthcare costs can be astronomical when it comes to cancer treatment. Given the steep costs of treatment, it is expected for a patient to learn about the cost estimates before starting the treatment process. Having this information would be immensely helpful for the patient to avoid debt and the debt related stress at the end of the treatment cycle. This would also help MD Anderson Cancer Center to improve patient experience and payer performance. Therefore, it is crucial for the Cancer Center to have a robust analytical tool to predict the stages and costs of treatment for a given set of conditions and patient attributes. By analyzing the various factors driving the costs and the attributes affecting the condition of the patient, the treatment costs can be predicted more accurately than the state of the art. This paper analyzes the various factors affecting the costs by measuring the variance of expenses at different stages of treatment. Further, the condition of the patient was modelled as a stochastic process to have an in-depth knowledge of the likelihood of the patient's condition getting better or worse given the stage at which the patient entered the system.

Table of Contents

List of Tables

List of Figures

Introduction

The University of Texas MD Anderson Cancer Center is one of the world's largest centers dedicated to the cause of cancer patient care, research education and prevention. [4] The Revenue Cycle Management team at the Cancer Center is primarily involved with performing cost analysis, measuring profitability and analyzing payer performance. The concern that they had with their analysis were the inaccuracies in predicting the cost for a new patient. The inaccuracies do not pose a major issue for insured patients, as their insurance will take care of the majority of the costs, but we have to keep in mind that copays can be very expensive too. However, the patients without any insurance have to face unexpected expenses which are not accounted for in their initial cost estimate.

United States has the most expensive healthcare system in the world. [3] The country doesn't have a nationwide system of health insurance, however insurance can be bought from a private marketplace or provided by the government. [2] Cancer treatment however, is very expensive and drags on for a very long period of time. Given that 75.1% of a staggering $5 billion worth of revenue earned by MD Anderson Cancer Center came from the patients [4] (both insured and self-pay), it is of importance for us to make the patient experience better by providing accurate estimations for their treatment and help them prepare better financially.

For our analysis we were provided with data sets for Melanoma, Breast Cancer and Gynecological Cancer. In the next sections, the paper discusses how factorial analyses were performed on certain categorical variables to identify features affecting the costs. This

analysis takes cues from the work done by Bertsimas Et al [1]. Further it discusses the process of obtaining the probability of the condition of a patient getting better or worse at different stages of treatment. Python packages like pandas and seaborn have been used extensively in this project for analyzing and visualizing the data respectively. The analysis sheds some light on the major factors driving up the costs and provides valuable insights on the treatment cycle of cancer patients.

Objectives

In order to understand the core factors driving the costs the initial objectives were to perform exploratory analysis into the dataset. The key objectives were to:

1. Find cost estimates for new patients over their period of stay, especially for Self-Pay[1] patients. These estimates were measured by expenses per month and cumulative expenses over the period of treatment.
2. Identify the important factors that affect the costs by analyzing the variance of costs split by the factor levels. For example, the costs were split by gender to see if there was a significant difference between the treatment costs for males and females.
3. Investigate the metastasis of cancer by observing the stochastic nature of the acuity levels of the patient over the period of treatment.

[1] Patients without an insurance cover for their treatment.

Revenue Cycle

Before diving into understanding the dataset and the analysis, it is of importance for us to understand how the Revenue Cycle works. Here "cycle" means the entire process of treatment starting from admission or pre-admission process till the final bill settlement. This process goes through various stages of treatment, accounting and settlement.

Once a patient is admitted for a particular treatment, a Patient Medical Record Number (Patient MRN) and a Hospital Account Record (HAR) is created for the patient. This account keeps track of all the expenses falling under that treatment cycle. A patient can have multiple HAR records based on the different treatments received. Each procedure performed on the patient is called an encounter and each encounter is a subset of a particular HAR. These encounters are identified by a Contact Serial Number (CSN). A bunch of CSNs come under the umbrella of a unique HAR; similarly, a bunch of HARs constitute a Patient MRN. The tree diagram of the hierarchy of the codes is given in figure 1**.**

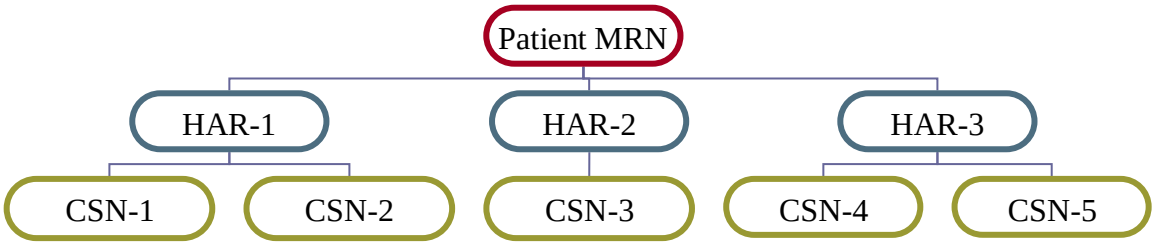

Figure 1: Schematic representation of cost clustering hierarchy.

The entire cycle is represented by figure 2. The diagram shows the revenue cycle starting from the patient's admission/pre-admission process till the point when the account is fully cleared and given a "$0 Account Balance" status.

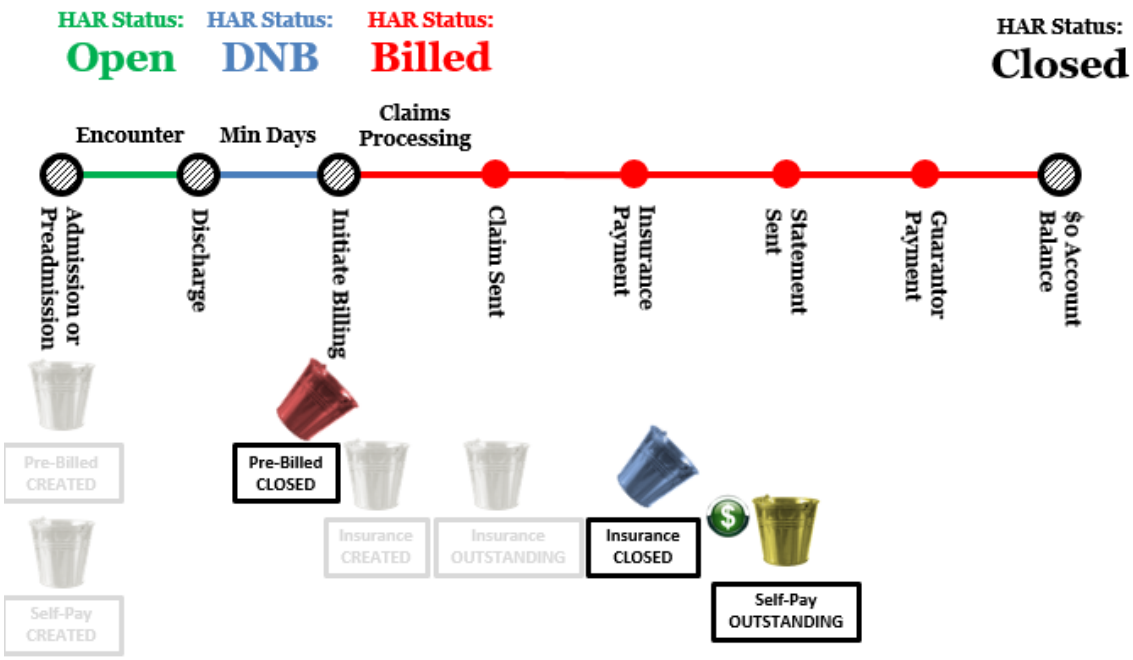

Figure 2: Schematic representation of revenue cycle. [2]

[2] Image Courtesy: The UT MD Anderson Cancer Center

Dataset

For our analysis, we obtained the datasets for patients diagnosed and treated for melanoma, breast cancer, and gynecologic cancer. The datasets consisted of costs incurred by the patients per procedure per day, along with additional details for identifying features impacting the costs starting from March of 2016 to March of 2018. Each dataset consisted of two sheets of information. One was the Hospital Billing (HB) sheet which was the records of costs for the material/facility use during the treatment. Here "material" means the treatments like chemotherapy, blood transfusions, injections etc. The second sheet was the Provider Billing (PB) sheet. This sheet provided the cost breakdowns for the service charges by various providers involved in the treatment process, e.g. nurses, surgeons etc.

The datasets obtained were in .xlsx format and were imported to python using pandas. Pandas is a powerful tool in Python which helps in easier cleansing of datasets and has optimized functions to slice and dice the data. Pandas' built in functions were used to convert the dataset into a dataframe[3] for easier usability in the code.

After parsing the data, it was observed that there were 32 columns in the HB dataset and some of those did not have any direct relation with the cost of the patient, but those factors were important for clustering the various branches of expenses. The cost clustering has not been performed in our analysis. Various branches of expenses refer to the tree diagram as given in figure 1. The different columns of importance were the "Patient MRN", "Gender", "Age", "Primary Payer", "Service Date", "Registration Date", "Procedure

[3] A data structure representing cases (rows), each of which consists of a number of observations or measurements (columns).

codes", "CPT/HCPCS codes", "Revenue Codes" and "Total charge". These fields have been described in detail below.

AGE

The age attribute makes intuitive sense as a factor for analysis, since the age vulnerability, required treatments and recovery rates will be different for patients in different age groups. For our initial analysis we checked the number of patients in each age group. It was observed that the patients in the age group between 55 and 65 were the most vulnerable to Melanoma and Gynecologic cancers. While the vulnerable age for breast cancer was observed to be between 40 and 60. Further, given that breast cancer is more predominant in females and gynecologic cancer affects females only, it is worthwhile to note that women have a marginally higher risk of developing gynecologic cancer in their 60s than that of breast cancer, going by the observations.

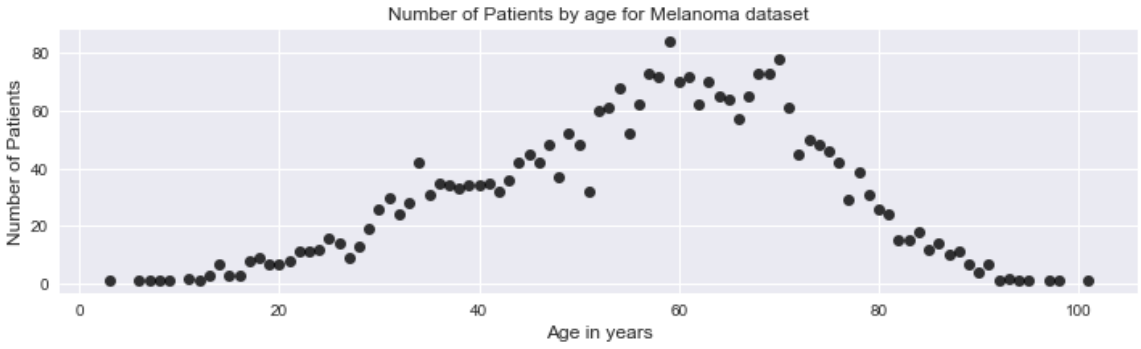

Figure 3: Number of patients by age for Melanoma dataset

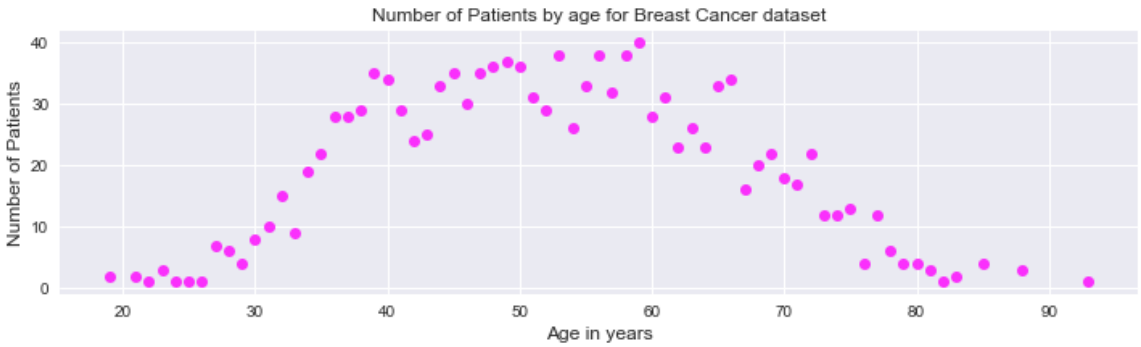

Figure 4: Number of patients by age for Breast Cancer dataset

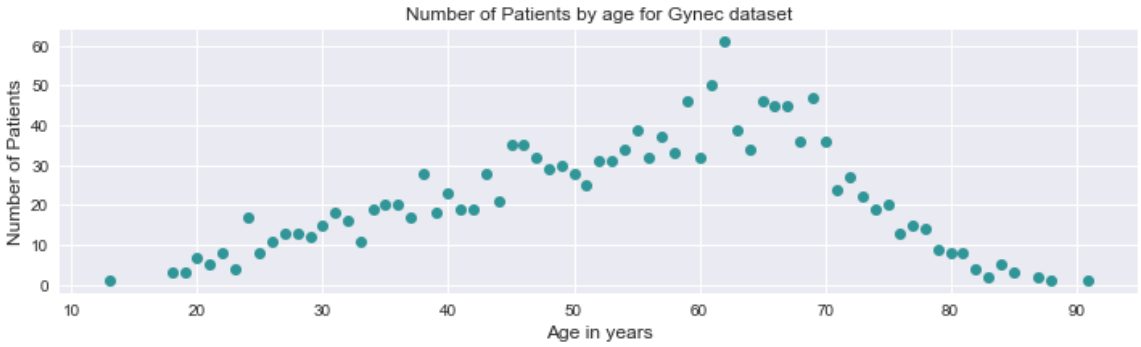

Figure 5: Number of patients by age for Gynecologic Cancer dataset

GENDER

Similar to age, gender was also taken into consideration to observe the differentiation in the treatment cost. But the effect of gender was not considered for the breast cancer and gynecologic cancer datasets. For example, in the gynecologic cancer dataset, out of the 3167 patients in the dataset, only 5 patients were Male. Similarly, we observed that only 45 out of 4562 patients in the Breast cancer dataset were males. Plotting

the total cost of treatment with gender as a factor would show a very high variability for males for the above-mentioned datasets. Therefore, including gender as a factor would not provide a robust insight to the associated costs for breast cancer and gynecologic cancer patients.

PRIMARY PAYER

The primary payer field identifies the source of payment for the cancer treatment. The patients paying out of pocket have been marked as "Self-Pay". If the insurance company is paying for the treatment, then the name of the particular insurance provider is mentioned in this column. We identified that a very small portion of the patients were Self-Pay. But that was the basic objective of our analysis. After looking more thoroughly into the data, it was observed that there was no difference between the procedure costs for the Self-Pay patients compared to the insured patients, and all patients were treated equally. This provided enough evidence to drop the payer type differentiation, as the goal of cost estimation can still be achieved without it.

SERVICE DATE

This variable records the date on which the patient received a particular service. This field was used to measure the stages of the expenses incurred over the length of treatment. This field also came in handy for the Breast Cancer and Gynecologic Cancer datasets as the minimum value in this column for each patient was used as the registration date.

REGISTRATION DATE

The registration date field represents the date on which a patient was registered on the MD Anderson system to receive treatment. This field gives us the starting point of treatment and helps us measure the relative time spent in treatment.

The calendar month and year of services are not related to cost of treatment. Rather, we want to know how many months have passed since the time a patient was admitted for treatment. The time of treatment here implies the time elapsed from the start of treatment till the current service date inn consideration. This way, we can analyze the data for all patients based on their time of registration irrespective of the date of admission.

Thus,

Time of treatment = Service Date – Registration Date

The Registration Date field was available for the Melanoma dataset. However, the same information was not available for the Breast Cancer and Gynecologic Cancer dataset. For those cases, we assumed that the earliest service date for a patient in the dataset as Registration Date.

For Gynecologic and Breast Cancer datasets,

Time of treatment = Service Date – Minimum of Service Date

CPT/HCPCS CODE

Current Procedural Terminology (CPT) or the Healthcare Common Procedure Coding System (HCPCS) codes are useful in identifying specific procedure types. For our

study, we noticed that chemotherapy is one of the major methods in cancer treatment. So the average amounts were estimated separately for chemo and non-chemo treatments.

The procedure codes starting with the letter "J" were identified as chemotherapy related. The patients with at least one CPT/HCPCS code starting with the letter J were marked as chemo and other patients were marked as non-chemo.

REVENUE CODE

The revenue code section provides us valuable information about the various procedure groups that drive up the costs. Analyzing the effect of procedures directly as a factor is not very effective in our analysis, since there are hundreds of different procedures involved in the treatment for the different patients. That is where the revenue code groupings data is useful to perform analysis over fewer factors.

TOTAL CHARGE

Each procedure performed on a patient on a single date has a cost associated with it. The procedure might be repeated multiple times on the same day of treatment. The Total Charge column accounts for the charge incurred for each procedure multiplied by the number of repetitions. The cost per procedure can be derived by dividing the total charge by the "qty" (Quantity[4]) column. Per procedure cost is not an important factor in our analysis so we won't be going into the details of this variable.

[4] The quantity column represents the repetition of a procedure on the same day of treatment.

Preprocessing

Before stepping into the factorial analysis, it was important for us to standardize the length of treatment for all the patients by setting the start of treatment to a base of zero (as discussed in the Service Date/Registration Date sections previously). A new column was added to the imported dataset to represent the number of months of treatment elapsed since the start of treatment. This column was named "nb_months". The snippet in figure 6 demonstrates the same.

	Reg Date	Service Date	nb_months
0	2016-03-29	2016-05-05	1
1	2016-03-29	2016-05-03	1
2	2016-03-29	2016-05-03	1
3	2016-03-29	2016-05-03	1
4	2016-03-29	2016-05-03	1
5	2016-03-29	2016-05-03	1

Figure 6: New column to account for the time elapsed in treatment in months

After adding the treatment time column, it was easy to combine the costs per month per patient. Other attributes like age, gender etc. were also preserved to be used later for factorial analysis. Next, to differentiate the patients on the basis of chemotherapy treatments the code (Appendix 1) iterated through the entire dataframe and identified the Patient MRNs with at least one CPT/HCPCS code starting with J and stored it in a separate list named "chemoMRN". A new column was created to identify the chemo/non-chemo patients. The default value was set to "non-chemo". The code (Appendix 1) iterated

through the dataframe again to check which Patient MRNs were available in the "chemoMRN" list and changed the value to "chemo" in the "chemo/non-chemo" column. A sample data snippet is provided in figure 7, where we observe the total charge amount per month by length of treatment for a patient who underwent chemotherapy.

Patient Gender	nb_months	Charge Amount (Total)	chemo/non-chemo
Female	0	15492.08	chemo
Female	1	18618.04	chemo
Female	2	9282.18	chemo
Female	3	15336.03	chemo

Figure 7: A snippet demonstrating a patient data for chemotherapy.[5]

[5] Patient MRN field has been omitted from this snippet to comply with The UT MD Anderson Cancer Center's guidelines.

Data Analysis

The analysis of the costs starts off with the melanoma dataset. Since this dataset has the gender differentiation, it would be of value to check that analysis first. Then we will be moving on to the Breast Cancer and Gynecologic cancer analyses.

MELANOMA DATASET

To have a clear idea about the variabilities and averages we are going to observe ahead, some initial probing is needed. The average cost of treatment over the time-elapsed with some information on the number of patients in the system gives us a good initial measure of costs and variability factors. Figure 8 helps us visualize the same. As we can observe from the graph, the costs shoot up steeply for the initial 6 months of treatment, after which the costs settle down relatively towards the later stages of treatment. Further, the number of patients decline rapidly after the 4^{th} month of treatment. Therefore, the variability in the costs will visibly increase as the number of observations thin out towards the later stages.

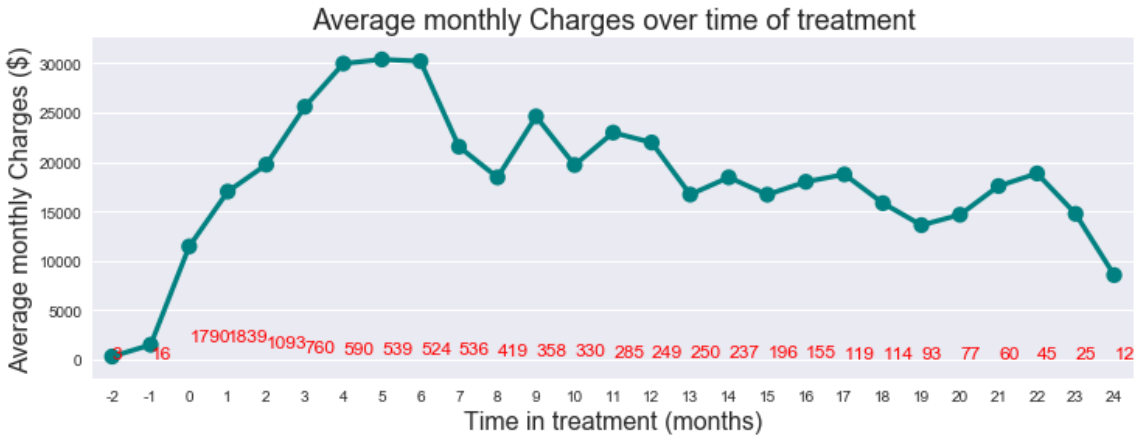

Figure 8: Average monthly charges for Melanoma patients over the period of treatment with number of patients marked in red.

Gender as a differentiating factor

Next the analysis for the dataset is shifted towards the differentiation levels, by observing the spread of the expenses per month for patient gender as the differentiating factor as shown in figure 8. The graphs show the total charges per month over the duration of treatment for HB and PB charges. The Provider (PB) charges are not very different for male and female patients over the period of treatment. However, it is clearly observed that the charges for male and female patients are very different from each other for the HB charges. As per table 1, the ratio of average charges for male vs female melanoma patients is roughly 1.45.

When we take a look at the cumulative charges for the patients over the period of treatment, we observe that the treatment costs are similar for male and female patients for the early stage of the treatment, however the costs start to diverge significantly after the 9th

month of treatment. Further, we notice that the costs seem to have a high variance towards the later stages of treatment. This variability is visibly higher due to fewer number of observations in that time frame and 90% confidence interval limit. The vertical bars represent the confidence interval for each of these graphs.

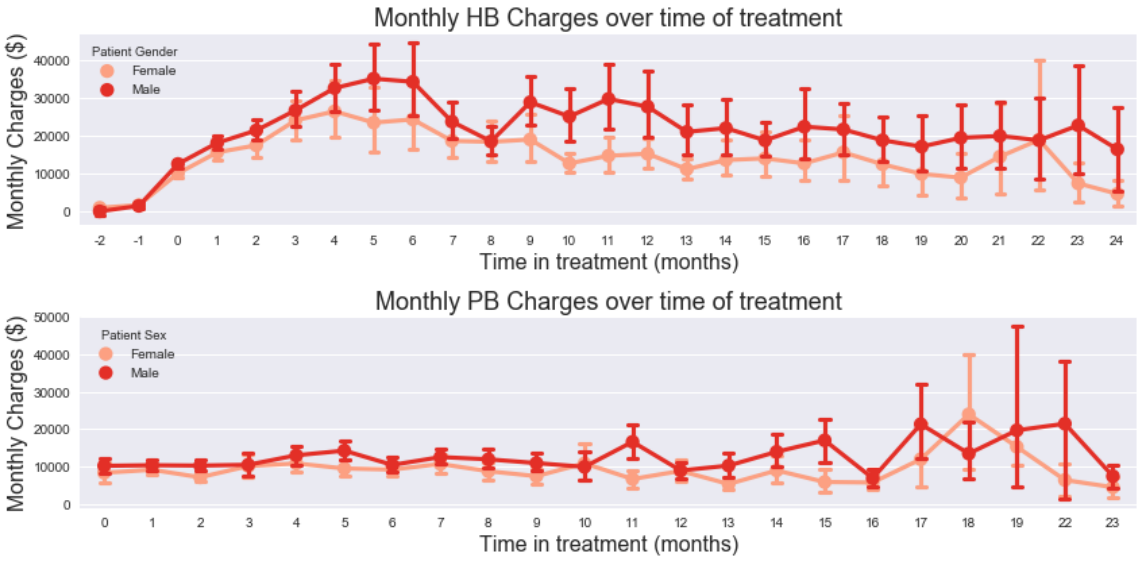

Figure 9: Average monthly charges for Melanoma patients over the period of treatment factored by gender

Gender	Count of Patients	Average Charges per month (in USD)
Female	1272	59958
Male	1565	87126

Table 1: Average charges in USD by gender for Melanoma patients approximated to the nearest whole number.

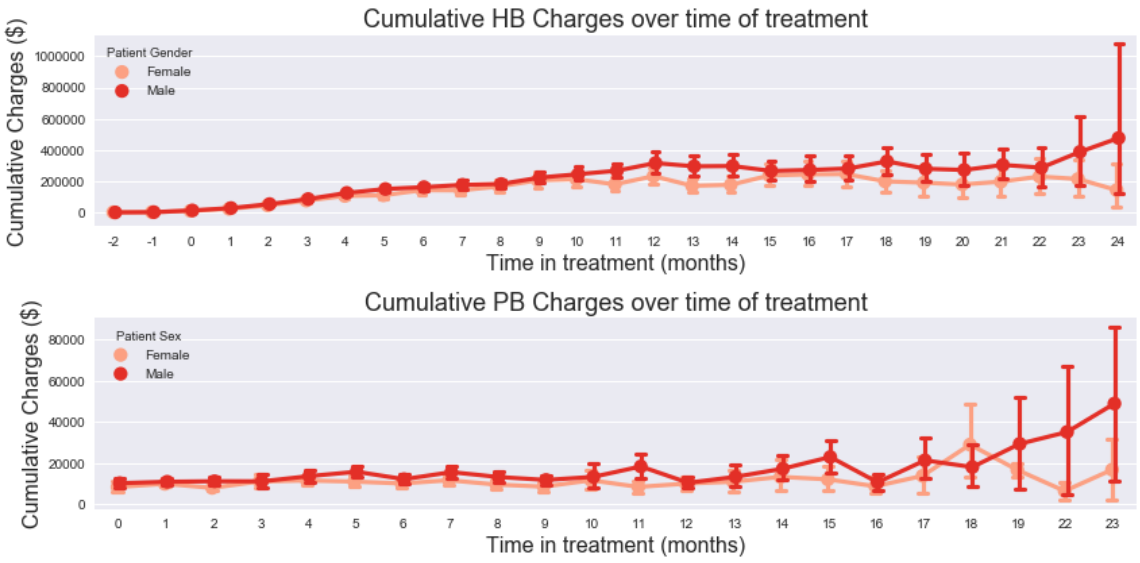

Figure 10: Cumulative charges for Melanoma patients over the period of treatment factored by gender

Age as a differentiating factor

Just like the differentiated the costs on the basis of gender, the expenses can be differentiated on the basis of age. For this representation the ages were grouped by a range of 20 years and a graph of the cumulative expenses was plotted for the HB and PB charges, as given by figure 11 below. Here we observe that the melanoma patients in the age range of 40-60 years are charged the highest for treatments. This might be an indication that the acuity levels for this age group might be higher compared to other age groups.

Further, the patients in the 0-20 years age group finish their treatment the earliest. The cost of patients in the 20-40 years age group has the highest variance.

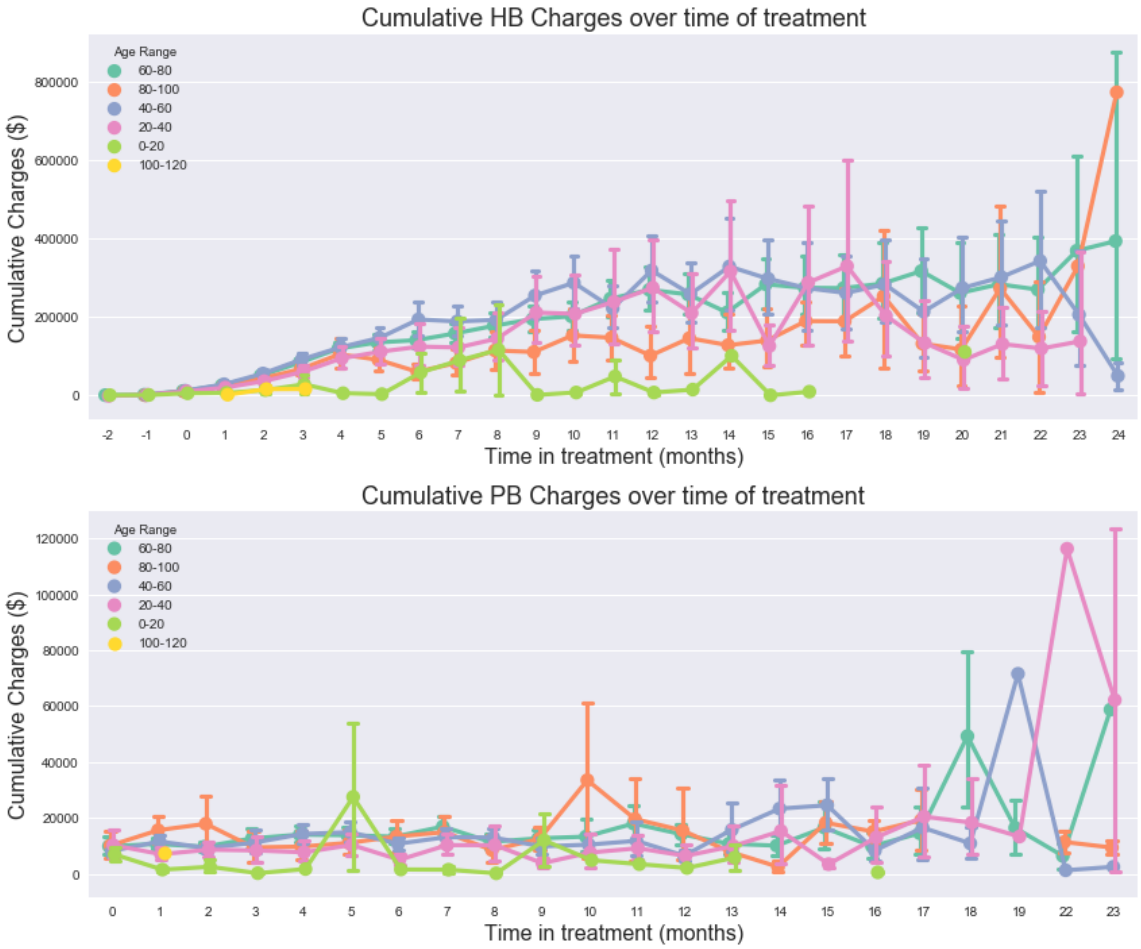

Figure 11: Cumulative charges for Melanoma patients over the period of treatment factored by age

Chemotherapy as a differentiating factor

Out of all the analyses we have performed, this differentiation gives us the most significant divergence in patient costs. For this analysis we have marked out the patients who underwent chemo therapy and who didn't. Figure 12 illustrates the difference in costs. As we observe here, the cost for patients undergoing chemotherapy is not even comparable to the costs of patients who do not. This provides us with useful insight towards cost estimation for non-chemo type patients. We can predict the costs for non-chemo patients

with a very high degree of accuracy. In this graph we see an anomaly in the starting values for the time in treatment. The negative values signify that the patient(s) got registered for treatment after a few preliminary tests at MD Anderson's facilities. Hence, although their services were recorded under their Patient MRN, but the registration date was after the initial services.

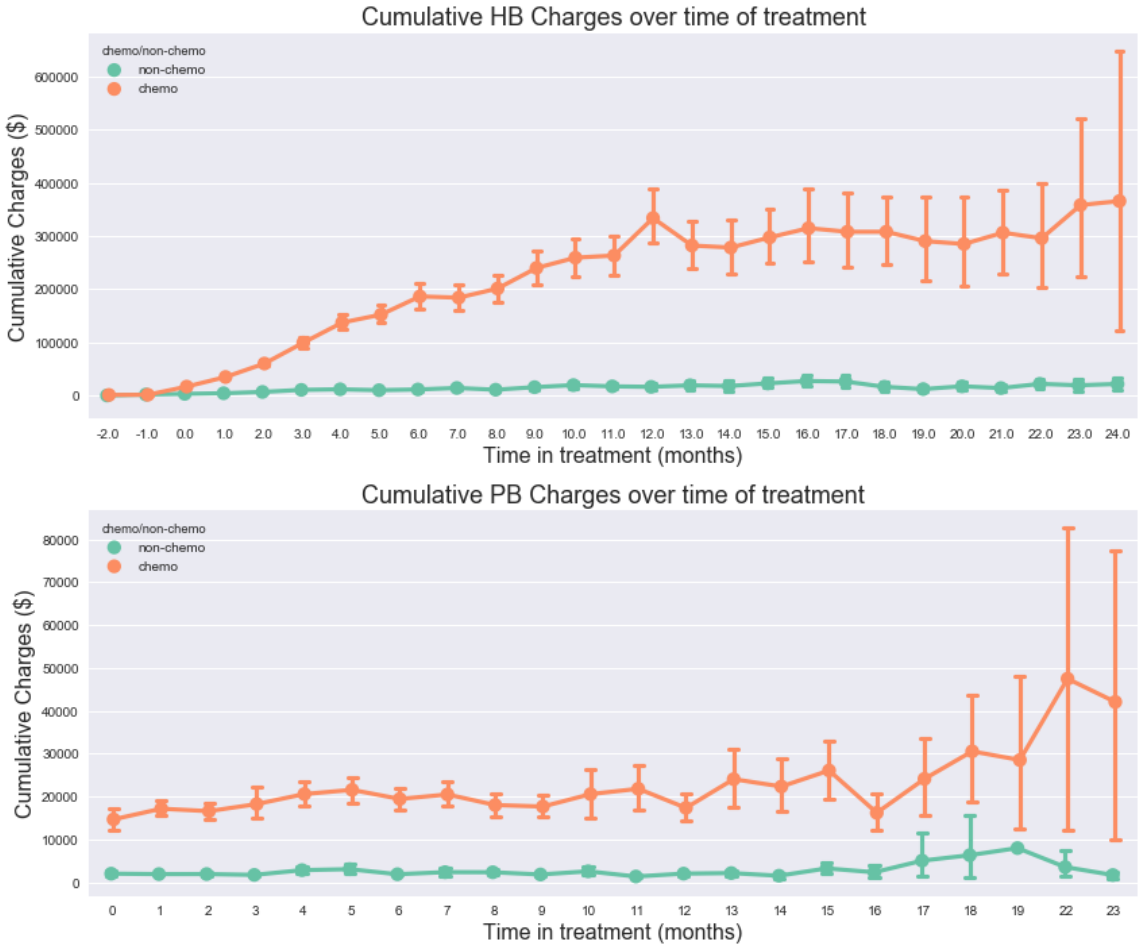

Figure 12: Cumulative charges for Melanoma patients over the period of treatment factored by chemo/non-chemo treatments

BREAST CANCER AND GYNECOLOGIC CANCER DATASETS

The analyses performed for both of these data sets are similar since gender is not a significant factor for either. Hence, we will only look into the effects of age and chemotherapy on the cumulative cost of treatment.

Age as a differentiating factor

By following a similar approach as the previous dataset, figures 13 and 14 illustrate the effect of age on breast cancer and gynecologic cancer treatment costs respectively. The graphs for HB and PB charges for breast cancer show a lot of similarity.

The average cost for patients is highest in the 20-40 years range, followed by 40-60, 60-80 and 80-100. This ordering applies to both the charges. Further, it sheds some light on the 0-20 range age group. As we can see, there aren't many instances of patients in that age group. Further a long-term treatment was not necessary in that case and the costs were very manageable.

Coming to the Gynecologic cancer data (figure 14), we observe that the average cost of treatment for women in the range of 20-40 years is relatively lower than other age ranges. Further the instances of this type of cancer showing up in the below 20 years age range is very rare, however the cost of treatment can be very high.

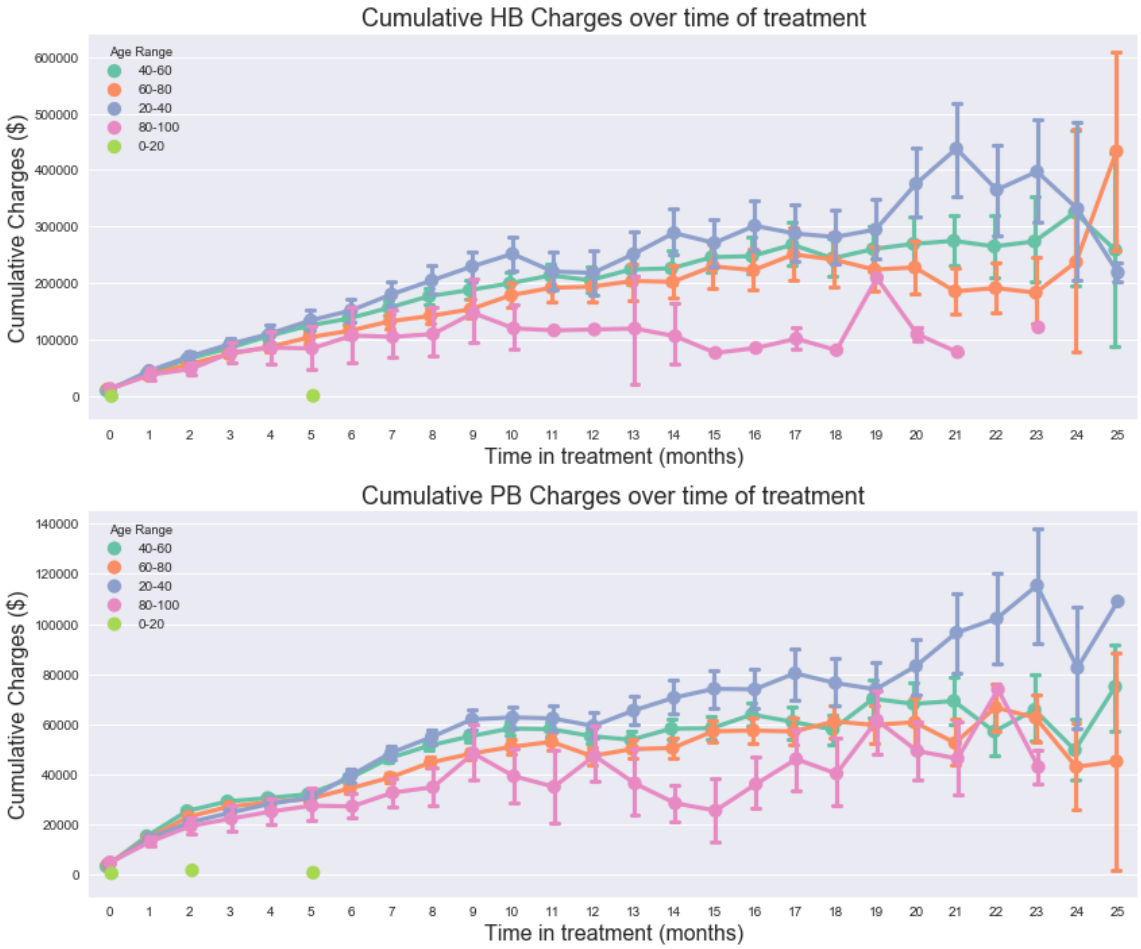

Figure 13: Cumulative charges for Breast Cancer patients over the period of treatment factored by age

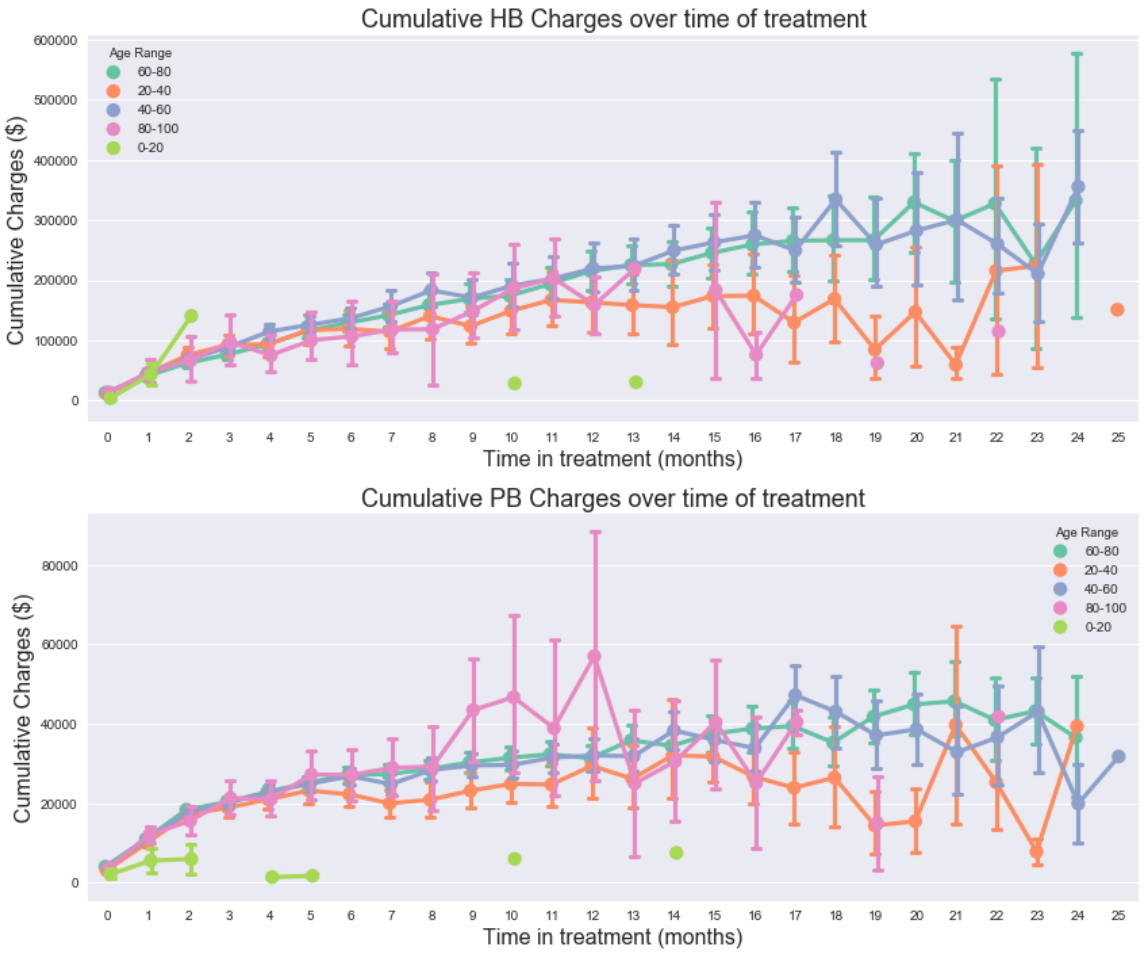

Figure 14: Cumulative charges for Gynecologic Cancer patients over the period of treatment factored by age

Chemotherapy as a differentiating factor

Figures 15 and 16 illustrate the effect of chemotherapy on the cumulative costs. For both the breast cancer and gynecologic cancer datasets, we observe similar patterns as we did for the melanoma dataset. The patients undergoing chemotherapy end up paying way higher in comparison to the non-chemo patients. Also, the cost for non-chemo patients can be estimated fairly accurately even for a long period of treatment.

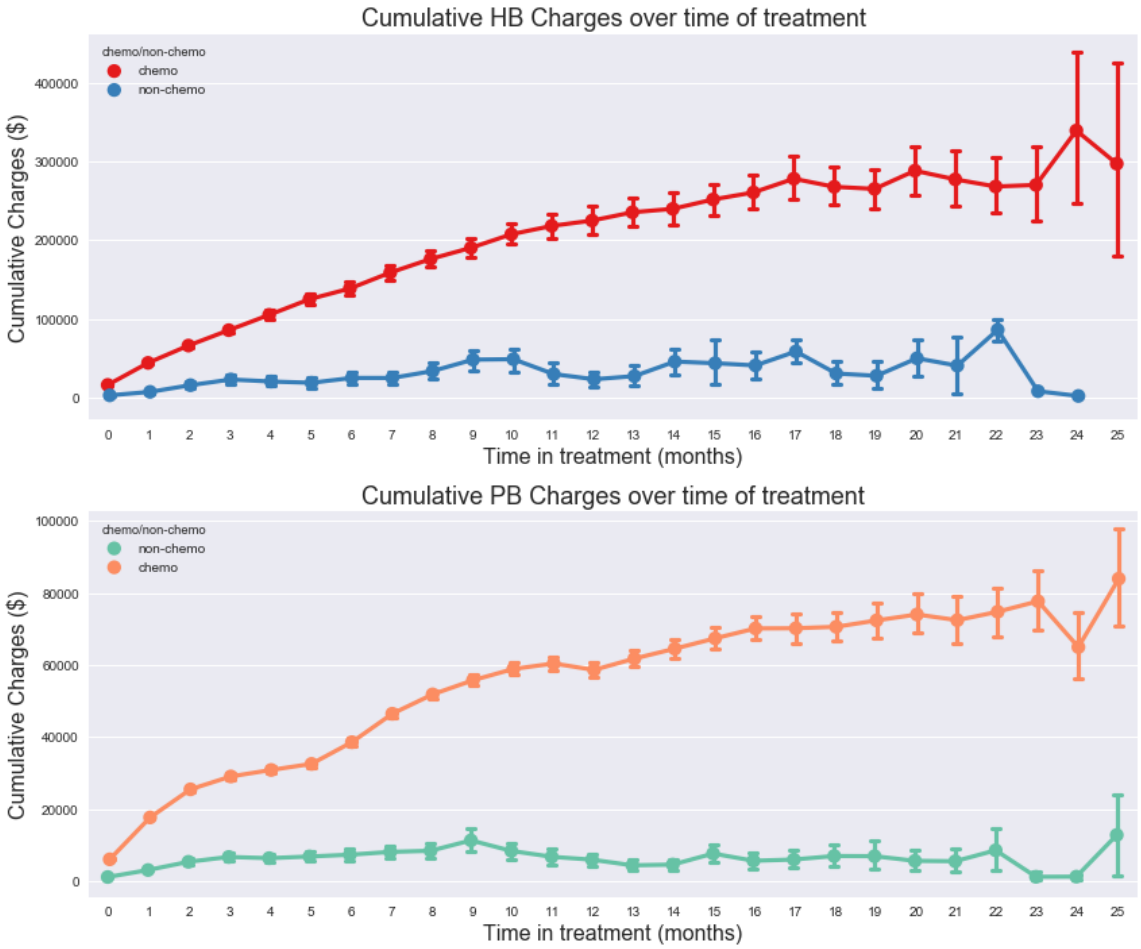

Figure 15: Cumulative charges for Breast Cancer patients over the period of treatment factored by chemotherapy

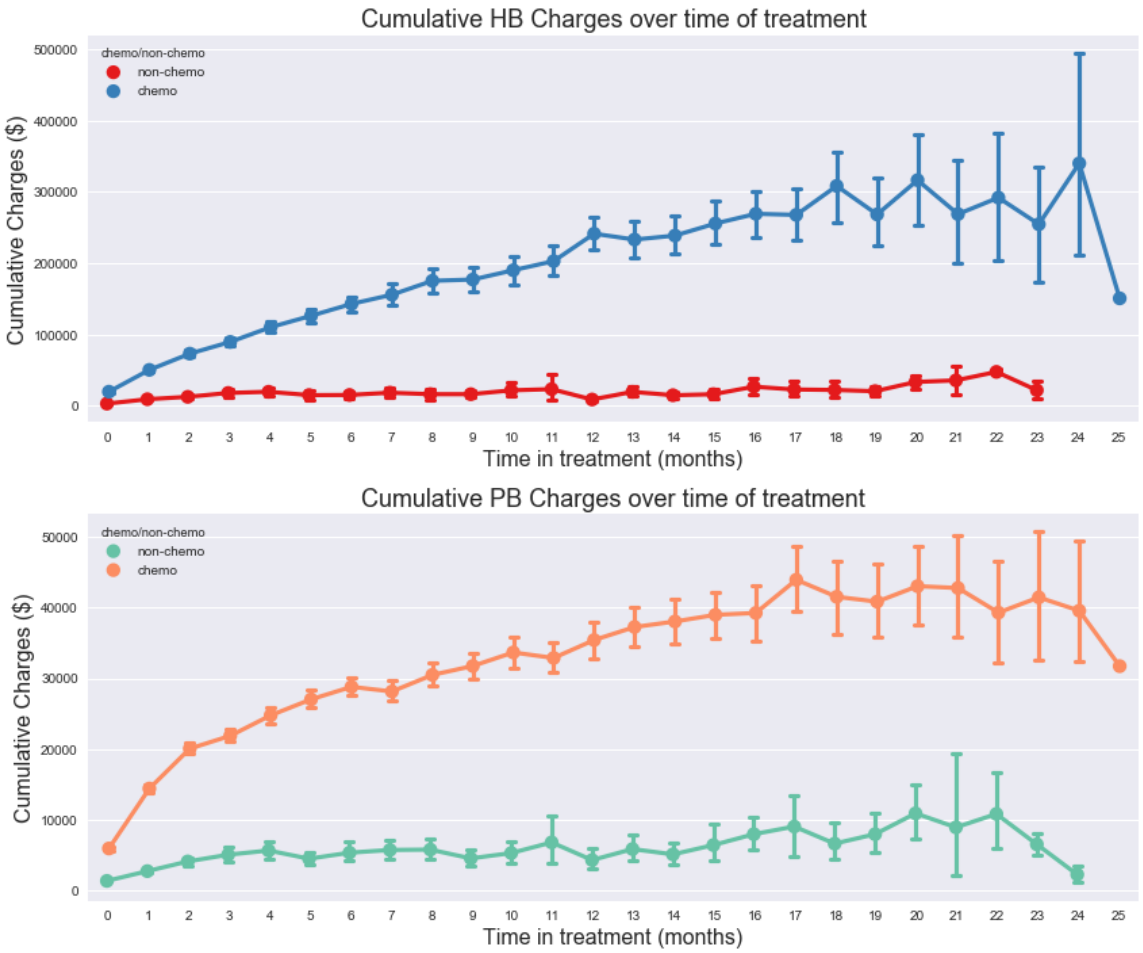

Figure 16: Cumulative charges for Gynecologic Cancer patients over the period of treatment factored by chemotherapy

Acuity Level Transition Rates

The information about the transition rates from one acuity level to another can play a significant role in predicting the cost of the treatment. The acuity level defines the severity of the patient condition and the level of treatment required to resolve the same. Level 1 is defined as a low acuity and level 5 as high acuity. The digits from 1 to 5 are derived from the unit's place number from the table in appendix 2. For the calculation to make some sense, we defined a value "0" which represents a state when a patient has not yet entered the system or the state which has not been recorded after the last service date for a particular patient.

The concept of a Discrete Time Markov Chain has been used here to create a transition matrix. The following steps were followed to obtain the matrices:

1. The "Patient MRN", "Service date" and "CPT/HCPCS code" columns were extracted from the dataset to create a new dataframe.
2. This new dataframe was sorted by service date to create a chronological order of records.
3. The code iterated over the dataframe to check if the CPT code matched with the ones provided in appendix 2. If it did then the last digit of the CPT code was extracted and appended to the list of last digits for a specific Patient MRN.
4. A dictionary of lists was created with Patient MRNs as keys and list of last digits as the value.
5. Next, as per the algorithm, the digit zero "0" was prefixed to each list to identify "outside of the MD Anderson system" status. Then all the lists were concatenated.
6. The algorithm then created another dictionary by storing the values of the acuity levels as keys along with the zero, and the list of probabilities of transitioning to

another digit as the values. For example, the key 1 might have the following probability values (dummy values used as a representation):

1: [0: 3.6%, 1: 0%, 2: 56%, 3: 22%, 4: 10%, 5: 8.4%]

In the above representation we can observe that the transition from acuity level 1 to 1 is zero, this is because we are not considering same state transitions as the information on the order of placement of acuity level in the CPT codes is unclear. Further, we notice that the probabilities sum up to 100%, which means that the patient will definitely transition to another state. To read the transition table, consider the digits from 0 to 5 written vertically as the acuity level at any given point of time and the horizontal digits 0 to 5 as the possible future acuity levels. "0" in the current time step represents the status of the person as "Outside of MD Anderson system", whereas the "0" in the future step represents an unknown state, since the dataset is not yet exhaustive about the full treatment cycle.

Tables 2, 3 and 4 show the transition rate probabilities for Melanoma, Breast Cancer and Gynecologic Cancer. The color schemes are picked to match the ribbon colors representing each cancer type.[6] Further the patchy nature of the colors provide a heat map of values. Higher probabilities have darker tones whereas lower probabilities fade towards white.

These probability values can provide crucial insights towards the effect of treatment and cost predictions. A cost estimate can be associated to an acuity level and the probability of transition between the different acuity levels can provide us with the basis to find a weighted average of the total cost.

[6] Black (used grey) for Melanoma, Pink for Breast Cancer and Teal for Gynecologic Cancer

Although it is theoretically possible to gauge the cost from these probability matrices, it's unclear how the acuity level might change in the future for the patients in the dataset. A longer-term dataset will be more valuable for such an analysis.

		Acuity level at the next time step					
		0	**1**	**2**	**3**	**4**	**5**
Acuity Level at the current time step	**0**	0.00%	1.40%	2.20%	55.00%	37.50%	3.90%
	1	5.40%	0.00%	3.10%	47.40%	39.70%	4.40%
	2	0.80%	5.80%	0.00%	69.90%	22.40%	1.10%
	3	7.50%	9.00%	9.80%	0.00%	68.20%	5.60%
	4	55.30%	2.20%	1.60%	39.10%	0.00%	1.90%
	5	43.00%	1.80%	1.60%	33.40%	20.10%	0.00%

Table 2: Acuity level transition rates for Melanoma patients.

		Acuity level at the next time step					
		0	**1**	**2**	**3**	**4**	**5**
Acuity Level at the current time step	**0**	0.00%	3.10%	3.80%	50.30%	26.50%	16.20%
	1	1.10%	0.00%	3.00%	49.70%	31.70%	14.60%
	2	1.30%	3.90%	0.00%	60.10%	26.00%	8.80%
	3	3.30%	9.00%	8.10%	0.00%	60.10%	19.60%
	4	9.50%	5.70%	4.10%	66.10%	0.00%	14.50%
	5	52.30%	2.70%	1.70%	26.30%	17.00%	0.00%

Table 3: Acuity level transition rates for Breast Cancer patients.

		Acuity level at the next time step					
		0	**1**	**2**	**3**	**4**	**5**
Acuity Level at the current time step	**0**	0.00%	1.40%	1.30%	44.30%	43.30%	9.70%
	1	1.60%	0.00%	3.00%	43.40%	42.70%	9.40%
	2	2.30%	5.50%	0.00%	49.90%	35.20%	7.10%
	3	7.00%	8.80%	5.20%	0.00%	61.10%	17.90%
	4	38.40%	4.60%	2.10%	45.60%	0.00%	9.30%
	5	46.40%	3.60%	1.40%	28.40%	20.10%	0.00%

Table 4: Acuity level transition rates for Gynecologic Cancer patients.

Results

The analysis provided us with some valuable insights regarding the various factors affecting the cost of treatment for the various kinds of cancer. We observed that each of the analyzed factors, i.e. age, gender and chemotherapy play a major role in differentiating the costs. Chemotherapy status showed the highest effect on the cost. Whereas age differentiation showed subtle but important differences in costs for each age range.

Coming to the transition rate probabilities, we observed that melanoma patients (figure 3) have the highest probability of entering the system at acuity level 3. Further, the probability of the condition worsening from level 2 to 3 and 3 to 4 are the highest in the matrix. Similar patterns were observed for Breast Cancer and Gynecologic Cancer acuity levels. But, due to the limitation on the length of the observations, it was hard to gauge the transition from the higher acuity levels back to lower ones. This is portrayed by the high transition rates from level 4 and 5 to level 0. Datasets over longer time periods might be more useful in these cases.

Based on these observations, we can approach towards building more comprehensive models to predict costs. The data collected from the patient before registering him/her on the system is crucial for us to predict the possible outcomes and the associated cost during the treatment. Having this information would help us build a more robust training set. Further, the categorical variables present in the dataset, e.g. diagnosis codes, procedure codes etc., need to be limited to fewer factor levels to fruitfully analyze their effects on the cost.

Appendix 1: Pseudo Codes

Pseudo Code for chemo/NonChemo:

```
START

SET ChemoMRN to 0
FOR each PatientMRN find RegDate = min(ServiceDate)

Time-of-Treatment = ServiceDate – RegDate

IF CPTcode starts with J THEN
     Append ChemoMRN
ELSE
     BREAK
END IF

GROUPBY PatientMRN, Gender and Time-of-Treatment for ChemoMRN

PLOT data by Cumulative cost and Average monthly cost

END
```

Pseudo Code for Acuity levels:

```
START

SET acuityMRN to 0
FOR each PatientMRN find RegDate = min(ServiceDate)

Time-of-Treatment = ServiceDate – RegDate

CONVERT PatientMRN to Object

IF CPTcode ends with acuity-level THEN
     Append acuityMRN
ELSE
     BREAK
END IF

GROUPBY PatientMRN, Gender and Time-of-Treatment for acuityMRN

PLOT data by Cumulative monthly prices and Average monthly prices

END
```

Appendix 2: Acuity Level codes

CPT Category	CPT HCPCS Code	
ESTABLISHED OUTPATIENTS	99211	Low Acuity
	99212	
	99213	
	99214	
	99215	High Acuity
NEW PATIENTS	99201	Low Acuity
	99202	
	99203	
	99204	
	99205	High Acuity
OUTPATIENT CONSULTS	99241	Low Acuity
	99242	
	99243	
	99244	
	99245	High Acuity

Table 5: Acuity levels by CPT codes

References

[1] Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., & Wang, G. (2008). Algorithmic prediction of health-care costs. Operations Research, 56(6), 1382-1392.

[2] Ridic, G., Gleason, S., & Ridic, O. (2012). Comparisons of Health Care Systems in the United States, Germany and Canada. Materia Socio-Medica, 24(2), 112–120. http://doi.org/10.5455/msm.2012.24.112-120

[3] Sarpel, U., Vladeck, B. C., Divino, C. M. & Klotman, P.E (2008). Fact and fiction: debunking myths in the US healthcare system. Annals of Surgery, 247(4), 563-569.

[4] www.mdanderson.org (2018). Quick Facts. Retrieved from http://www.mdanderson.org/documents/about-md-anderson/about-us/facts-and-history/quick-facts.pdf.